

Towards a Books Data Commons for AI Training



April 2024



1. Introduction¹

While the field of artificial intelligence research and technology has a long history, broad public attention grew over the last year in light of the wide availability of new generative AI systems, including large language models (LLMs) like GPT-4, Claude, and LLaMA-2. These tools are developed using machine learning and other techniques that analyze large datasets of written text, and they are capable of generating text in response to a user's prompts.

While many large language models rely on website text for training, books have also played an important role in developing and improving AI systems. Despite the widespread use of e-books and growth of sales in that market, books remain difficult for researchers and entrepreneurs to access at scale in digital form for the purposes of training AI.

In 2023, multiple news publications reported on the availability and use of a dataset of books called "Books3" to train LLMs.² The Books3 dataset contains text from over 170,000 books, which are a mix of in-copyright and out-of-copyright works. It is believed to have been originally sourced from a website that was not authorized to distribute all of the works contained in the dataset. In lawsuits brought against OpenAI, Microsoft, Meta, and Bloomberg related to their LLMs, the use of Books3 as training data was specifically cited.³

The Books3 controversy highlights a critical question at the heart of generative AI: what role do books play in training AI models, and how might digitized books be made widely accessible for the purposes of training AI? What dataset of books could be constructed and under what circumstances?

In February 2024, Creative Commons, Open Future and Proteus Strategies convened a series of workshops to investigate the concept of a responsibly designed, broadly accessible dataset of digitized books to be used in training AI models. Conducted under the Chatham House Rule, we set out to ask if there is a possible future in which a "books data commons for AI training" might exist, and what such a commons might look like. The workshops brought together practitioners on the front lines of building next-generation AI models, as well as legal and policy scholars with expertise in the copyright and licensing challenges surrounding digitized books. Our goal was also to bridge the perspective of stewards of

¹ Authored by Alek Tarkowski and Paul Keller (Open Future), Derek Slater and Betsy Masiello (Proteus Strategies) in collaboration with Creative Commons. We are grateful to participants in the workshops, including Luis Villa, Tidelift and openml.fyi; Jonathan Band; Peter Brantley, UC Davis; Aaron Gokaslan, Cornell; Lila Bailey, Internet Archive; Jennifer Vinopal, HathiTrust Digital Library; Jennie Rose Halperin, Library Futures/NYU Engelberg Center, Nicholas P. Garcia, Public Knowledge; Sayeed Choudhury, Erik Stallman, UC Berkeley School of Law. The paper represents the views of the authors, however, and should not be attributed to the workshop as a whole. All mistakes or errors are the authors'.

² See e.g. Knibbs, Kate. "The Battle over Books3 Could Change AI Forever." *Wired*, 4 Sept. 2023, www.wired.com/story/battle-over-books3/.

³ For key documents in these cases, see the helpful compendium at "Master List of Lawsuits v. AI, ChatGPT, OpenAI, Microsoft, Meta, Midjourney & Other AI Cos." *Chat GPT Is Eating the World*, 27 Dec. 2023, chatgptiseatingtheworld.com/2023/12/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos. See also "Fair Use Week 2024: Day Two with Guest Expert Brandon Butler." *Fair Use Week*, sites.harvard.edu/fair-use-week/2024/02/26/fair-use-week-2024-day-two-with-guest-expert-brandon-butler/. Accessed 20 Mar. 2024 (arguing that use of this dataset is not consequential for the fair use analysis).

content repositories, like libraries, with that of AI developers. A “books data commons” needs to be both responsibly managed, and useful for developers of AI models.

We use “commons” here in the sense of a resource that is broadly shared and accessible, and thus obviates the need for each individual actor to acquire, digitize, and format their own corpus of books for AI training. This resource could be collectively and intentionally managed, though we do not mean to select a particular form of governance in this paper.⁴

This paper is descriptive, rather than prescriptive, mapping possible paths to building a books data commons as defined above and key questions relevant to developers, repositories, and other stakeholders, building on our workshop discussions. We first explain why books matter for AI training and how broader access could be beneficial. We then summarize two tracks that might be considered for developing such a resource, highlighting existing projects that help foreground both the potential and challenges. Finally, we present several key design choices, and next steps that could advance further development of this approach.⁵

⁴ In this way, we do not use “commons” in the narrow sense of permissively licensed. What’s more, this resource could also be governed as more of a data “trust,” and, indeed, we discuss extensively the work of HathiTrust as a relevant project in this domain. However, our use of the word “commons” is not meant to preclude this or other arrangements.

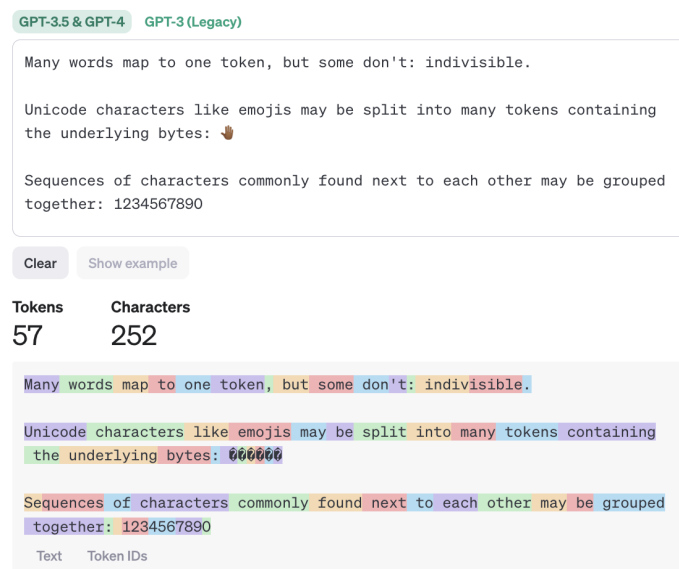
⁵ There are, of course, a range of other types of texts that are not on the web and/or not digital at all - e.g., periodicals, journals, government documents. These are out of scope for this paper, but also worthy of further analysis.

2. Basics of AI Training and Technical Challenges of Including Books

It's critical to understand that LLMs are not trained on text "as is" – meaning that the model is not digesting the text in a way humans would, front to back. The text does not represent a copy of the original text in its original form. Instead, the text is processed in smaller chunks of text, which are then shuffled and "tokenized," as we explain further below.

One way to conceptualize the chunking, shuffling and tokenizing process is to imagine a 900 page book, which has 400,000 words. To feed into an AI model, the book will first be cut into manageable chunks of text that represent up to several thousand tokens; such a process might result in around 50 "chunks" of text. Each of those chunks will contain long sections of narrative content; however, the chunks themselves will then be randomized, and fed into the AI model out of sequence from each other; the first chunk may be text from Chapters 9 and 10, while the initial text in Chapter 1 may be in the 30th chunk. Within these chunks, the text itself will be understood by the model as tokens.

In the example below, 252 characters of human-readable text are shown in tokenized form as 57 distinct tokens, the relationships between which then form the basis of building an AI model. The illustration shows a block of human-readable text as it would be tokenized for AI training; different colors are used in this visualization merely to differentiate one token from another within the string of text. As the visualization makes clear, not all of the tokens directly correspond to a single word; tokens merely represent characters that often appear together in the training data.⁶



The screenshot shows the OpenAI tokenizer interface. At the top, it says "GPT-3.5 & GPT-4" and "GPT-3 (Legacy)". The input text is: "Many words map to one token, but some don't: indivisible. Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌 Sequences of characters commonly found next to each other may be grouped together: 1234567890". Below the input, there are "Clear" and "Show example" buttons. A table shows the results: Tokens: 57, Characters: 252. Below the table, the text is visualized with colored boxes representing tokens. The text is: "Many words map to one token, but some don't: indivisible. Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌 Sequences of characters commonly found next to each other may be grouped together: 1234567890". At the bottom, there are "Text" and "Token IDs" tabs.

⁶ OpenAI's Tokenizer tool at <https://platform.openai.com/tokenizer> explains how ChatGPT uses tokens and provides a tool to visualize examples. As noted on their site, the tokenization process is different for every model, this is merely an illustrative example. The visual below represents an example of how OpenAI's ChatGPT creates tokens from English text.

Tokens do not typically represent words, but instead often represent subword tokens. For example the word “incompetence” may be broken into three tokens: “in-,” “competent,” and “-ence.” This approach to tokenization enables representation of grammar and word variations, effectively allowing a high degree of language generalizability.⁷

In recent years, LLM research has successfully been able to scale up models by pre-training on a large number of tokens. In turn, this has allowed a higher degree of language generalizability in the resulting model. For example, OpenAI’s ChatGPT trained on hundreds of billions of tokens, allowing it to model language in a very general way. The resulting models can then be fine-tuned for specific tasks using training data representing a particular corpus, such as software code.⁸

⁷ McKinsey provides an overview of the different types of tokens that may be used by AI models. McKinsey. “What Is Tokenization? | McKinsey.” *Mckinsey.com*, 2023, www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-tokenization.

⁸ There are certain technical challenges in using books in AI training as well, given the nature of the format. First, one must address whether a book is already in digital form. For the vast majority of books, that is not the case. One first needs to digitize the book, and convert it to a digital text file using optical character recognition (OCR), or use a born-digital version (although we return to specific limitations on that approach below). Second, once a book is in digital text form, it must be converted into a text format that is suitable for AI training. Text conversion tools transfer the content of books into complete text files, which is akin to the type of conversion that must be done between a Microsoft Word or Adobe PDF file format and a simple .txt format. This conversion is generally not adequate for the purpose of AI training; researchers have found that post-processing is required to ensure these text files are properly formatted for the purposes of tokenization. For example, when building the dataset known as The Pile, researchers had to modify an existing epub-to-text converter tool to ensure that document structure across chapters was preserved to match the table of contents, that tables of data were correctly rendered, to convert numbered lists from digitally legible lists of “1\.” to “1.”, and to replace unicode punctuation with ascii punctuation. See Discussion in 4.3.2 in Bandy, Jack, and Nicholas Vincent. *Addressing “Documentation Debt” in Machine Learning Research: A Retrospective Datasheet for BookCorpus*. 2021, <https://arxiv.org/pdf/2105.05241.pdf>. and C.16 of The Pile documentation in Gao, Leo, et al. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, <https://arxiv.org/pdf/2101.00027.pdf>.

3. Why Books are Important to Training AI

Despite the proliferation of online content and some speculating that books would simply die out with the advent of the Internet,⁹ books remain a critical vehicle for disseminating knowledge. The more scientists study how books can impact people, the less surprising this is. Our brains have been shown to interact with longform books in meaningful ways: we develop bigger vocabularies when we read books; we develop more empathy when we read literary fiction; and connectivity between different regions of our brain increases when we read.¹⁰

In that light, it might be unsurprising that books are important for training AI models. A broadly accessible books dataset could be useful not only for building LLMs, but also for many other types of AI research and development.

Performance and Quality

The performance and versatility of an AI model can significantly depend on whether the training corpus includes books or not. Books are uniquely valuable for AI training due to several characteristics.

- **Length:** Books tend to represent longer-form content, and fiction books, in particular, represent long-form narrative. An AI trained on this longer-form, narrative type of content is able to make connections over a longer context, so instead of putting words together to form a single sentence, the AI becomes more able to string concepts together into a coherent whole; even after a book is divided into many “chunks” before the process of tokenization, that will still provide long stretches of text that are longer than the average web page. While Web documents, for instance, tend to be longer than a single sentence, they are not typically hundreds of pages long like a book.
- **Quality:** The qualities of the training data impact the outputs a tool can produce. Consider an LLM trained on gibberish; it can learn the patterns of that gibberish and, in turn, produce related gibberish, but will not be very useful for writing an argument or a story, for instance. In contrast, training an LLM on books with well-constructed arguments or crafted stories could serve those purposes. While “well-constructed” and “crafted” are necessarily subjective, the traditional role of editors and the publishing process can provide a useful indicator for the quality of writing inside of books. What’s more, metadata for books – information such as the title, author and year of publication – is often more comprehensive than metadata for information

⁹“the novel, too, as we know it, has come to its end” – “The End of Books.” *Archive.nytimes.com*, 21 June 1992, archive.nytimes.com/www.nytimes.com/books/98/09/27/specials/coover-end.html. Accessed 27 Aug. 2021.

¹⁰ Stanborough, Rebecca Joy. “Benefits of Reading Books: For Your Physical and Mental Health.” *Healthline*, 15 Oct. 2019, www.healthline.com/health/benefits-of-reading-books#prevents-cognitive-decline.

found on the web, and this additional information can help contextualize the provenance and veracity of information.

- **Breadth, Diversity, and Mitigating Bias:** Books can serve a critical role in ensuring AI models are inclusive of a broad range of topics and categories that may be under-represented in other content. For all that the Internet has generated an explosion in human creativity and information sharing, it generally represents only a few decades of information and a small portion of the world’s creative population. A books dataset, by comparison, is capable of representing centuries of human knowledge. As a result such a dataset can help ensure AI systems behavior is based on centuries of historical information from modern books. It can help ensure broad geographic and linguistic diversity. What’s more, the greater breadth and diversity of high-quality content help mitigate challenges around bias and misinformation. Using a more diverse pool of training data can help support the production of a model and outputs of the model that are more representative of that diversity. Books can be useful in evaluation datasets to test existing models for memorization capabilities, which can help prevent unintended reproduction of existing works. Of course, this is all contingent on actual composition of the corpus; in order to have the benefits described, the books would need to be curated and included with characteristics like time, geographic and linguistic diversity.
- **Other Modalities:** Finally, books do not just contain text, they often contain images and captions of those images. As such, they can be an important training source for multi-modal LLMs, which can receive and generate data in media other than text.

Lowering Barriers to Entry & Facilitating Competition

Broad access to books for AI training is critical to ensure powerful AI models are not concentrated in the hands of only a few companies. Access to training data, in general, has been cited as a potential competitive concern¹¹ in the AI field because of the performance benefits to be gained by training on larger and larger datasets. But this competitive wedge is even more acute when we look specifically at access to book datasets.

The largest technology companies building commercial AI models have the resources and capacity to mass digitize books for AI training. Google has scanned 40 million books, many of which came from digitization partnerships they formed with libraries. They may already use some or all of these books to train their AI systems.¹² It’s unclear to what extent other companies already have acquired books for AI training (for instance, whether Amazon’s existing licenses with publishers or self-published authors may permit such uses);

¹¹ See e.g. Trendacosta, Katherine and Doctorow, Cory. “AI Art Generators and the Online Image Market.” *Electronic Frontier Foundation*, 3 Apr. 2023, www.eff.org/deeplinks/2023/04/ai-art-generators-and-online-image-market; Narechania, Tejas N., and Sitaraman, Ganesh. “An Antimonopoly Approach to Governing Artificial Intelligence.” *SSRN Electronic Journal*, 2023, cdn.vanderbilt.edu/vu-URL/wp-content/uploads/sites/412/2023/10/09151452/Policy-Brief-2023.10.08-.pdf, <https://doi.org/10.2139/ssrn.4597080>. Accessed 25 Feb. 2024.

¹² See white paper for Google’s Gemini models <https://arxiv.org/pdf/2312.11805.pdf> – “Gemini models are trained on a dataset that is both multimodal and multilingual. Our pretraining dataset uses data from web documents, books, and code, and includes image, audio, and video data.”

regardless, comparable efforts to Google's would cost many hundreds of millions of dollars.¹³

Independent researchers, entrepreneurs, and most other businesses and organizations are unlikely to have the resources required to digitally scan millions of books nor purchase licenses to digitized books in ways that could unlock the benefits described above. Ensuring greater competition and innovation in this space will require making this type of data available to upstarts and other entities with limited resources. A well-designed and appropriately governed digital books commons is one way to do that.

¹³ "By 2004, Google had started scanning. In just over a decade, after making deals with Michigan, Harvard, Stanford, Oxford, the New York Public Library, and dozens of other library systems, the company, outpacing Page's prediction, had scanned about 25 million books. It cost them an estimated \$400 million. It was a feat not just of technology but of logistics." Somers, James. "Torching the Modern-Day Library of Alexandria." *The Atlantic*, 20 Apr. 2017, www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/.

4. Copyright, Licensing, & Access to Books for Training

Even if books can be acquired, digitized, and made technically useful for AI training, the development of a books data commons would necessarily need to navigate and comply with copyright law.

Out-of-Copyright Books: A minority of books are old enough to be in the public domain and out of copyright, and an AI developer could use them in training without securing any copyright permission. In the United States, all books published or released before 1929 are in the public domain. While use of these books provides maximal certainty for the AI developer to train on, it is worth noting that the status of whether a book is in the public domain can be difficult to determine.¹⁴ For instance, books released between 1929 and 1963 in the U.S. are out of copyright if they were not subject to a copyright renewal; however, data on copyright renewals is not easily accessible.

What's more, copyright definitions and term lengths vary among countries. Even if a work is in the public domain in the US, it may not be in other countries.¹⁵ Countries generally use the life of the last living author + "x" years to determine the term of copyright protection. For most countries, "x" is either 50 years (the minimum required by the Berne Convention) or 70 years (this is the case for all member states of the European Union and for all works published in the U.S. after 1978). This approach makes it difficult to determine copyright terms with certainty because it requires information about the date of death of each author, which is often not readily available.

In-Copyright Books: The vast majority of books are in copyright, and, insofar as the training process requires making a copy of the book, the use in AI training may implicate copyright law. Our workshop covered three possible paths for incorporating such works.

Direct licensing

One could directly license books from rightsholders. There may be some publishers who are willing to license their works for this purpose, but it is hard to determine the scale of such access, and, in any event, there are significant limits on this approach. Along with the challenge (and expense) of reaching agreements with relevant rightsholders, there is also the practical difficulty of simply identifying and finding the rightsholder that one must negotiate

¹⁴ For a sense of the complexity, see e.g. Melissa Levine, Richard C. Adler. *Finding the Public Domain: Copyright Review Management System Toolkit*. 2016, quod.lib.umich.edu/c/crmstoolkit/14616082.0001.001. Accessed 20 Mar. 2024.; Kopel, Matthew. "LibGuides: Copyright at Cornell Libraries: Copyright Term and the Public Domain." guides.library.cornell.edu/copyright/publicdomain; Mannapperuma, Menesha, et al. *Is It in the Public Domain? A HANDBOOK for EVALUATING the COPYRIGHT STATUS of a WORK CREATED in the UNITED STATES*. 1923.

¹⁵ See e.g. Moody, Glyn. "Project Gutenberg Blocks Access in Germany to All Its Public Domain Books because of Local Copyright Claim on 18 of Them." *Techdirt*, 7 Mar. 2018, www.techdirt.com/2018/03/07/project-gutenberg-blocks-access-germany-to-all-public-domain-books-because-local-copyright-claim-18-them/. Accessed 20 Mar. 2024.

with. The vast majority of in-copyright books are out-of-print or out-of-commerce, and most are not actively managed by their rightsholders. There is no official registry of copyrighted works and their owners, and existing datasets can be incomplete or erroneous.¹⁶

As a result, there may be no way to license the vast majority of in-copyright books, especially those that have or have had limited commercial value.¹⁷ Put differently, the barrier to using most books is not simply to pay publishers; even if one had significant financial resources, licensing would not enable access to most works.

Permissively licensed works

There are books that have been permissively licensed in an easily identifiable way, such as works placed under Creative Commons (CC) licenses. Such works explicitly allow particular uses of works subject to various responsibilities (e.g., requiring attribution by the user in their follow-on use).

While such works could be candidates for inclusion in a books data commons, their inclusion depends on whether the license's terms can be complied with in the context of AI training. For instance, in the context of CC licensed works, there are requirements for proper attribution across all licenses (the CC tools Public Domain Dedication (CC0) and Public Domain Mark (PDM) are not licenses and do not require attribution).¹⁸

¹⁶ See e.g. Heald, Paul J. "How Copyright Makes Books and Music Disappear (and How Secondary Liability Rules Help Resurrect Old Songs)." Illinois Program in Law, Behavior and Social Science Paper No. LBSS14-07 Illinois Public Law Research Paper No. 13-54 <https://doi.org/10.2139/ssrn.2290181>. Accessed 4 Jan. 2020, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2290181; Rosen, Rebecca J. "Why Are so Few Books from the 20th Century Available as Ebooks?" *The Atlantic*, 18 Mar. 2014, www.theatlantic.com/business/archive/2014/03/why-are-so-few-books-from-the-20th-century-available-as-ebooks/284486/. See also "Google Book Search Settlement and Access to Out of Print Books." *Google Public Policy Blog*, publicpolicy.googleblog.com/2009/06/google-book-search-settlement-and.html. Accessed 20 Mar. 2024 (discussing this issue in the context of the failed class-action settlement between Google, the Authors Guild, and the Association of American Publishers). Google's final brief in the settlement proceedings notes the "prohibitive transaction costs of identifying and locating individual Rightsholders of these largely older, out-of-print books" — see this brief at https://web.archive.org/web/20130112060651/http://thepublicindex.org/docs/amended_settlement/google_final_approval_support.pdf. The Authors Guild and Association of American Publishers also justified the settlement's terms in light of the fact that "the transaction costs involved in finding copyright owners and clearing the rights are too high"; while they argued that most works are not truly "orphans," they note that total transaction costs as a whole (including, for example, determining whether the author or publisher holds the rights and then negotiating rates) are so high as to block uses of out-of-print works anyway — see this brief at https://web.archive.org/web/20130112060213/http://thepublicindex.org/docs/amended_settlement/Supplemental_memorandum_of_law.pdf.

¹⁷ In the EU, the 2019 Copyright Directive introduced specific provisions on the "use of out-of-commerce works and other subject matter by cultural heritage institutions" (Articles 8-11 CDSMD). These provisions allow cultural heritage institutions to "make available, for non-commercial purposes, out-of-commerce works or other subject matter permanently in their collections". The limitation to non-commercial purposes means that works made available under these provisions would be of limited use in building a books data commons.

¹⁸ For one assessment of the difficulties of complying with the CC licenses in this context, to the extent they are applicable, see Lee, K., A. Feder Cooper, & Grimmelmann, J. (2023). Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain. Forthcoming, *Journal of the Copyright Society* 2024. <https://doi.org/10.2139/ssrn.4523551>.

Reliance on Copyright Limitations and Exceptions

Even if a book is in copyright, it's possible that copying books for AI training may be covered by existing limitations and exceptions to copyright law in particular jurisdictions. For example:

- In the United States, many argue using existing works to train generative AI is “fair use,” consistent with existing law and legal precedents.¹⁹ This is the subject of a number of currently active court cases, and different actors and tools may yield different results, as fair use is applied case-by-case using a flexible balancing test.
- In the European Union, there are explicit exceptions in the law for “text and data mining” uses of in-copyright works, both for non-commercial research and for commercial purposes. However, for commercial uses and for users outside of research and heritage institutions, they must respect the rights of rightsholders who choose to “reserve their rights” (i.e., opt-out of allowing text and data mining) via machine readable mechanisms.²⁰ The exception also requires that users have “lawful access” to the works.
- Finally, Japan provides a specific text and data mining exception, without any comparable opt-out requirement for commercial uses as is embedded in EU law.²¹

While exceptions that allow AI training exist in several other countries, such as Singapore and Israel, most countries do not provide exceptions that appear to permit AI training. Even where potentially available, as in the United States, legal uncertainty and risk create a hurdle for anyone building a books commons.²²

¹⁹ See e.g. Comments from Sprigman, Samuelson, Sag to Copyright Office, October 2023, at <https://www.regulations.gov/comment/COLC-2023-0006-10299> as well as many other submissions to the US copyright office; see also Advocacy, Katherine Klosek, Director of Information Policy and Federal Relations, Association of Research Libraries (ARL), and Marjory S. Blumenthal, Senior Policy Fellow, American Library Association (ALA) Office of Public Policy and. “Training Generative AI Models on Copyrighted Works Is Fair Use.” *Association of Research Libraries*, 23 Jan. 2024, www.arl.org/blog/training-generative-ai-models-on-copyrighted-works-is-fair-use/.

²⁰ See Articles 3 and 4 of the EU's Directive on Copyright and Related Rights in the Digital Single Market — <https://eur-lex.europa.eu/eli/dir/2019/790/oj>.

²¹ Japan clarified its laws in 2018 to make clear that this type of use is permitted — see discussion in Testimony of Matthew Sag, July 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-07-12_pm_-_testimony_-_sag.pdf, see also Fiil-Flynn, S. et al. (2022) *Legal reform to enhance global text and Data Mining Research, Science*. Available at: <https://www.science.org/doi/10.1126/science.add6124> (Accessed: 28 Sept. 2023).

²² See supra note 22. See also Jonathan Band, *Copyright Implications of the Relationship between Generative Artificial Intelligence and Text and Data Mining* | Infojustice. infojustice.org/archives/45509. In addition, for an in-depth look at the cross-border legal challenges involved see: *Wrapping up Our NEH-Funded Project to Help Text and Data Mining Researchers Navigate Cross-Border Legal and Ethical Issues*. 2 Oct. 2023, buildingltdm.org/2023/10/02/wrapping-up-our-neh-funded-project-to-help-text-and-data-mining-researchers-navigate-cross-border-legal-and-ethical-issues/. Accessed 20 Mar. 2024.

It is also important to note two other issues that can affect the application of limitations and exceptions, in particular, their application to e-books.

The first important limitation is that almost every digital book published today comes with a set of contractual terms that restrict what users can do with it. In many cases, those terms will explicitly restrict text data mining or AI uses of the content, meaning that even where copyright law allows for reuse (for example, under fair use), publishers by contract can impose restrictions anyway. In the United States, those contract terms are generally thought to override the applicability of fair use or other limitations and exceptions.²³ Other jurisdictions, such as those in the EU, provide that certain limitations and exceptions cannot be contractually overridden, though experience to date varies with how those anti-contractual override protections work in practice.²⁴

The second limitation is the widespread adoption of “anti-circumvention” rules in copyright laws and the interplay of these with a choice to rely on copyright limitations and exceptions. Digital books sold by major publishers are generally encumbered with “digital rights management” (DRM) that limits how someone can use the digital file. For instance, DRM can limit the ability to make a copy of the book, or even screenshot or excerpt from it, among other things. Anti-circumvention laws restrict someone's ability to evade these technical restrictions, even if it is for an ultimately lawful use.

What this means for our purposes is that even if one acquires a digital book from, for example, Amazon, and it is lawful under copyright law to use that book in AI training, it can still generally be unlawful to circumvent the DRM to do so, outside narrow exceptions.²⁵ Thus, the ability to use in-copyright books encumbered by DRM – that is, most all books sold by major publishers – is generally limited.²⁶

Practically, using in-copyright books to build a books commons for AI training – while relying on copyright's limitations and exceptions – requires turning a physical book into digital form, or otherwise engaging in the laborious process of manually re-creating a book's text (i.e., re-typing the full text of the book) without circumventing the technical restrictions themselves.

²³ See Hansen, Dave. “Fair Use Week 2023: How to Evade Fair Use in Two Easy Steps.” *Authors Alliance*, 23 Feb. 2023, www.authorsalliance.org/2023/02/23/fair-use-week-2023-how-to-evade-fair-use-in-two-easy-steps/. Accessed 20 Mar. 2024.

²⁴ See Band, Jonathan. “Protecting User Rights against Contract Override.” *Joint PIJIP/TLS Research Paper Series*, 1 May 2023, digitalcommons.wcl.american.edu/research/97/. Accessed 20 Mar. 2024.

²⁵ In the U.S. the Copyright Office has recognized the importance of allowing particular exceptions for researchers engaged in text and data mining. See their rulemaking in 2021 <https://www.federalregister.gov/documents/2021/10/28/2021-23311/exemption-to-prohibition-on-circumvention-of-copyright-protection-systems-for-access-control>. These rules are reviewed triennially and are currently under review, with submissions suggesting both contraction and expansion; see the Authors' Alliance comments in January 2024 <https://www.authorsalliance.org/2024/01/29/authors-alliance-submits-long-form-comment-to-copyright-office-in-support-of-petition-to-expand-existing-text-and-data-mining-exemption/>. It is possible that one could argue for these exceptions to be expanded, and then work to renew that exception every three years. The EU's text and data mining exception may also limit use of DRM to impede data mining, but only for particular covered research and heritage institutions; commercial and other users are not covered, however.

²⁶ Note that CC licenses forbid use of DRM – but that doesn't address most all books sold by publishers.

5. Examining approaches to building a books data commons

There are many possible permutations for building a books data commons. To structure our exploration, we focused on two particular tracks, discussed below. We chose these tracks mindful of the above legal issues, and because there are already existence proofs that help to illuminate tradeoffs, challenges and potential paths forward for each.

5a. Public domain and permissively licensed books

Existing Project Example²⁷: The Pile v2

In 2020, the nonprofit research group EleutherAI constructed and released The Pile – a large, diverse, open dataset for AI training. EleutherAI developed it not only to support their own training of LLMs, but also to lower the barriers for others.²⁸

Along with data drawn from the web at large, The Pile included books from three datasets. The first dataset was the Books3 corpus referenced at the outset of this paper. The second and third books datasets were smaller: BookCorpus2, which is a collection of 17,868 books by otherwise unpublished authors; and a 28,752 books in the public domain and published prior to 1919, drawn from a volunteer effort to digitize public domain works called Project Gutenberg.

As the awareness about The Pile dataset grew, certain rightsholders began sending copyright notices to have the dataset taken down from various websites.

Despite the takedown requests, the importance of books to EleutherAI and the broader community’s AI research remained. In hoping to forge a path forward EleutherAI announced in 2024 that they would create a new version of the dataset, which they will call The Pile v2.²⁹ Among other things, v2 would “have many more books than the original Pile had, for example, and more diverse representation of non-academic non-fiction domains.” At the same time, it would only seek to include public domain books and permissively licensed content. As before, this corpus focuses on English language books.

²⁷ This is an illustrative example, and there are also other projects of this ilk. For instance, see the Common Corpus project, which includes an array of public domain books from a number of countries, at <https://huggingface.co/blog/Pclanglais/common-corpus>; see also https://huggingface.co/datasets/storytracer/internet_archive_books_en (“This dataset contains more than 650,000 English public domain books (~ 61 billion words) which were digitized by the Internet Archive and cataloged as part of the Open Library project.”)

²⁸ See Gao et al, supra note 8.

²⁹ Goldman, Sharon. “One of the World’s Largest AI Training Datasets Is About to Get Bigger and ‘Substantially Better.’” *VentureBeat*, 11 Jan. 2024, venturebeat.com/ai/one-of-the-worlds-largest-ai-training-datasets-is-about-to-get-bigger-and-substantially-better/. Accessed 20 Mar. 2024.

Implications of the The Overall Approach

Stepping back from The Pile v2 specifically, or any particular existing collection of books or dataset built on their basis, we want to understand the implications of relying on public domain works and expressly licensed works in building a books commons.

The benefits are relatively straightforward. Both categories, by definition come with express permission to use the books in AI training. The cost of acquiring the books for this use may be effectively zero or close to it, when considering public domain and “openly” licensed books that allow redistribution and that have already been digitized.

But this approach comes with some clear limitations. First, as noted above, for many books in the public domain, their status as such is not always clear. And with respect to permissively licensed books, it is not always clear whether and how to comply with the license obligations in this context.

Setting aside those challenges, the simple fact is that relying on public domain and existing permissively licensed books would limit the quantity and diversity of data available for training, impacting performance along different dimensions. Only a small fraction of books ever published fall into this category, and the corpus of books in this category is likely to be skewed heavily towards older public domain books. This skew would, in turn, impact the content available for AI training.³⁰ For instance, relying on books from before 1929 would not only incorporate outdated language patterns, but also a range of biases and misconceptions about race and gender, among other things. Efforts could be made to get people to permissively license more material – a book drive for permissive licensing, so to speak; this approach would still not encompass most books, at least when it comes to past works.³¹

5b. Limitations & Exceptions

Existing Project Example: HathiTrust Research Center (HTRC)

The HathiTrust Research Center provides researchers with the ability to perform computational analysis across millions of books. While it is not suited specifically for AI training, it is an existence proof for what such a resource might look like.

³⁰ For instance, AI researchers note that the recently released Common Corpus dataset is an “invaluable resource” but “comes with limitations. A lot of public domain data is antiquated—in the US, for example, copyright protection usually lasts over seventy years from the death of the author—so this type of dataset won’t be able to ground an AI model in current affairs or, say, how to spin up a blog post using current slang” and the “dataset is tiny.” Thus, while it is possible to train an AI model on the data, those models will have more limited utility on some dimensions than current frontier models trained on a broader array of data. See Knibbs, Kate, *Here’s Proof You Can Train an AI Model Without Slurping Copyrighted Content | WIRED*. (2024, March 20), at <https://www.wired.com/story/proof-you-can-train-ai-without-slurping-copyrighted-content/>.

³¹ Our workshop discussion did note that some widely available datasets for AI training have also pursued more direct licensing agreements. For instance, the SILO LLM was created by working with scientific journal publishers to make works available for both download and AI training. While this might be viable in the context of particular, narrow classes of works, the barriers to efficient licensing mentioned above would remain a problem for any broader efforts. See Min, Sewon, et al. “SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore.” *ArXiv (Cornell University)*, 8 Aug. 2023, <https://doi.org/10.48550/arxiv.2308.04430>. Accessed 14 Dec. 2023.

It is also an example predicated on copyright's limitations and exceptions – in this case, on U.S. fair use. While the Authors Guild filed a copyright infringement suit against HathiTrust, federal courts in 2012 and 2014 ruled that HathiTrust's use of books was fair use.³²

A nonprofit founded in 2008, HathiTrust grew out of a partnership among major US university libraries and today is “an international community of research libraries committed to the long-term curation and availability of the cultural record.”³³ It started in what it calls the “early days of mass digitization” – that is, at a time when it started to become economical to take existing physical artifacts in libraries and turn them into digital files at a large scale.

The founding members of HathiTrust were among the initial partners for Google's Book Search product, which allows people to search across and view small snippets of text from in-copyright books³⁴ and read full copies of public domain books scanned from libraries' collections. The libraries provided Google with books from their collections, Google would then scan the books for use in Book Search, and return to the libraries a digital copy for their own uses. These uses included setting up HathiTrust not only to ensure long-term preservation of the digital books and their metadata, but also to facilitate other uses, including full text search of books and accessibility for people with print disabilities. In separate court cases, both Google and HathiTrust's uses of the books were deemed consistent with copyright law.

The uses most relevant to this paper are those enabled by what HathiTrust refers to today as the Research Center. The Center grew in part out of a research discipline called “digital humanities,” which, among other things, seeks to use computational resources or other digital technologies to analyze information and contribute to the study of literature, media, history, and other areas. For instance, imagine you want to understand how a given term (e.g., “war on drugs”) became used; one might seek to analyze when the term was first used and how often it was used over time by analyzing a vast quantity of sources, searching out the term's use. The insight here is that there is much to be learned not just from reading or otherwise consuming specific material, but also from “non-consumptive research,” or “research in which computational analysis is performed on one or more volumes (textual or image objects)” to derive other sorts of insights. AI training is a type of non-consumptive use.

Today, the Center “[s]upports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.” It includes over 18 million books in over 400 languages from the HathiTrust Digital Library collection. Roughly 58% of the corpus is in copyright. HathiTrust notes that, while this corpus is large, it has limitations in terms of its representation across subject matter, language, geography, and other dimensions. In terms of subject matter, the corpus is skewed towards humanities (64.9%) and social sciences (14.3%). In terms of language, 51% of the books are in English,

³² *Authors Guild v. HathiTrust*, 902 F.Supp.2d 445 (SDNY October 10, 2012) and *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).

³³ See <https://www.hathitrust.org/member-libraries/member-list/> – the membership is principally US institutions, and most of the non-US members are from English speaking countries or institutions that use English as the primary language of operations.

³⁴ This functionality is limited to scanned books provided by library partners in the US.

German is the next-largest language represented at 9%, and is followed by a long-tail of languages by representation.

In order to enable these uses, HathiTrust has invested in technical solutions to prevent possible misuse. To some extent, they manage this by limiting who gets access to the Center, and limiting access to specific features to researchers at member institutions. HathiTrust has also put in place various security controls on both the physical storage of the digitized books and the network access to those files. The primary uses of the data through the Research Center includes access to an extracted features set and access to the complete corpus “data capsule,” which is a virtual machine running on the Center’s servers. The data capsule allows users to conduct non-consumptive research with the data, but it limits the types of outputs allowed in order to prevent users from obtaining full content of in-copyright works. The measures taken include physical security controls on the data centers housing the information, as well as restrictions via network access and encryption of backup tapes. In the finding that HathiTrust use was a fair use and thus rejecting a lawsuit brought by the Authors Guild, the Court noted the importance of these controls.³⁵

Today, the Center’s tools are not suitable for AI training, in that they don’t allow the specific types of technical manipulation of underlying text necessary to train an AI. Nevertheless, the Center demonstrates that building a books data commons for computational analysis is possible, and in turn points to the possibility of creating such a resource for AI training.³⁶

Implications of Overall Approach

By relying on existing limitations and exceptions in copyright law, the number of books one could include in the corpus of a books data commons is far greater and more diverse. Of course, a bigger dataset doesn’t necessarily mean a higher quality dataset for all uses of AI models; as HathiTrust shows, even a multimillion book corpus can skew in various directions. Still, dataset size generally remains significant to an LLM’s performance – the more text one can train on, or rather the more tokens for training the model, the better, at least along a number of performance metrics.³⁷

While holding the potential for a broader and more diverse dataset, a key limitation in pursuing this approach is that it is only feasible where relevant copyright limitations and exceptions exist. Even then, legal uncertainty means that going down this path is likely to generate, at a minimum, expensive and time-consuming litigation and regulatory

³⁵ This is explained explicitly in the appeals court’s decision: *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).

³⁶ HathiTrust has also made available some data derived from books, such as the Extracted Features set: “HTRC releases research datasets to facilitate text analysis using the HathiTrust Digital Library. While copyright-protected texts are not available for download from HathiTrust, fruitful research can still be performed on the basis of non-consumptive analysis of transformative datasets, such as in HTRC’s flagship Extracted Features Dataset, which includes features extracted from full-text volumes. These features include volume-level metadata, page-level metadata, part-of-speech-tagged tokens, and token counts.” <https://analytics.hathitrust.org/datasets#top>.

³⁷ See Testimony of Chris Callison-Burch, July 2023, <https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf> (“As the amount of training data increases, AI systems’ capabilities for language understanding and their other skills improve.”); Brown, Tom, et al. *Language Models Are Few-Shot Learners*. 22 July 2020, at <https://arxiv.org/pdf/2005.14165.pdf> (“we find that performance scales very smoothly with model size”).

engagement. And, at least in the U.S., it could generate billions of dollars in damages if the specific design choices and technical constraints are not adequate to justify a finding of fair use.

This sort of books dataset could be built by expanding use of in-copyright books that have already been digitized from existing libraries and other sources. Specifically, workshop participants mentioned that the Internet Archive, HathiTrust, and Google as entities that have digitized books and could repurpose their use to build a books commons, although challenges with using these datasets were noted. The Internet Archive is in the midst of litigation brought by book publishers for its program for lending digital books; while not directly relevant to the issue of AI training using their corpus of books, this sort of litigation creates a chilling effect on organizations seeking to make new uses of these digitized books. Meanwhile, Google encumbered HathiTrust's digital copies with certain contractual restrictions, which would need to be addressed to develop a books dataset for AI training, and Google itself is unlikely to share its own copies while it provides them a competitive advantage.

Perhaps as a matter of public policy, these existing copies could be made more freely available. For instance, to ensure robust competition around AI and advance other public interests, policymakers could remove legal obstacles to the sharing of digitized book files for use in AI training. Alternatively, policymakers could go further and affirmatively compel sharing access to these digital book files for AI training.

It's possible that there could be a new mass digitization initiative, turning physical books into new digital scans. At least in theory, one could try to replicate the existing corpora of HathiTrust, for example, without Google's contractual limitations. At the same time, such an effort would take many years, and it seems unlikely that many libraries would want to go to the trouble to have their collections digitized a second time. Moreover, while new scans may provide some incremental benefit over use of existing ones (e.g., by using the most modern digitization and OCR tools and thus improving accuracy), there is no inherent social value to making every entity that wants to do or allow AI training invest in their own redundant scanning.

A new digitization effort could target works that have not been yet digitized. This may be particularly useful given that previous book digitization efforts, and the Google Books project in particular, have focused heavily (though not exclusively) on libraries in English-speaking countries. Additional digitization efforts might make more sense for books in those languages that have not yet been digitized at a meaningful scale. Any new digitization effort might therefore start with a mapping of the extent to which a books corpus in a given language has been digitized.

6. *Cross-cutting design questions*

The workshops briefly touched on several cross-cutting design questions. While most relevant for approaches that depend on limitations and exceptions, considerations of these questions may be relevant across both tracks.

Would authors, publishers, and other relevant rightsholders and creators have any ability to exclude their works?

One of the greatest sources of controversy in this area is the extent to which rightsholders of copyrighted works, as well as the original creators of such works (e.g., book authors in this context), should be able to prevent use of their works for AI training.

While a system that required affirmative “opt-in” consent would limit utility significantly (as discussed above in the context of directly licensing works), a system that allowed some forms of “opt-out” could still be quite useful to some types of AI development. In the context of use cases like development of LLMs, the performance impact may not be so significant. Since most in-copyright books are not actively managed, the majority of books would remain in the corpus by default. The performance of LLMs can still be improved across various dimensions without including, for example, the most famous writers or those who continue to commercially exploit their works and may choose to exercise an opt-out. Perhaps the potential for licensing relationships (and revenue) may induce some rightsholders to come forward and begin actively managing their works. In such a case, uses that do require a license may once again become more feasible once the rightsholder can be reached.

Workshop participants discussed different types of opt-outs that could be built. For example, opt-outs could be thought of not in blanket terms, but only as applied to certain uses, for example to commercial uses of the corpus, but not research uses. This could build on or mirror the approach that the EU has taken in its text and data mining exceptions to copyright.³⁸ Opt-outs might be more granular, by focusing on allowing or forbidding particular uses or other categories of users, given that rights holders have many different sets of preferences.

Another question is about *who* can opt-out particular works from the dataset. This could solely be an option for copyright holders, although authors might be allowed to exercise an opt-out for their books even if they don’t hold the copyrights. This might create challenges if the author and rightsholder disagree about whether to opt a particular book out of the corpus. Another related issue is that individual books, such as anthologies, may comprise works created (and rights held) by many different entities. The images in a book may have come from third-party sources, for instance, or a compendium of poetry might involve many

³⁸ In fact, as noted above, to the extent an AI model developer intends for their model to abide by the EU’s legal regime, they will have to abide by such opt-outs, at least if they are engaged in text and data mining for commercial uses and/or are users outside of the covered set of research and heritage institutions. A books data commons may incorporate opt-outs in particular to serve such EU-focused AI developers.

different rightsholders and authors. Managing opt-outs for so many different interests within one book may get overly complicated very fast.

In any event, creating an opt-out system will need some ways of authenticating whether someone has the relevant authority to make choices about inclusion of a work.

Who would get to use the books data commons? For what?

A commons might be made publicly available to all, as has been done with datasets like The Pile. Another possible design choice is to restrict access only to authorized users and to enforce particular responsibilities or obligations in return for authorization. Three particular dimensions of permitted uses and users came up in our discussions:

- **Defining and ensuring acceptable and ethical use:** Participants discussed to what extent restrictions should be put on use of the resource. In the case of HathiTrust, acceptable use is implicitly ensured by limiting access to researchers from member institutions; other forms of “gated access” are possible, allowing access only to certain types of users and for certain uses.³⁹ One can imagine more fine-grained mechanisms, based on a review of the purpose for which datasets are used. This imagined resource could become a useful lever to demand responsible development and use of AI; alongside “sticks” like legal penalties, this would be a “carrot” that could incentivize good behavior. At the same time, drawing the lines around, let alone enforcing, “good behavior” would constitute a significant challenge.
- **Charging for use to support sustainability of the training corpus itself:** While wanting to ensure broad access to this resource, it is important to consider economic sustainability, including support for continuing to update the resource with new works and appropriate tooling for AI training. Requiring some form of payment to use the resource could support sustainability, perhaps with different requirements for different types of users (e.g., differentiating between non-commercial and commercial users, or high-volume, well-resourced users and others).⁴⁰
- **Ensuring benefits of AI are broadly shared, including with book authors or publishers:** The creation of a training resource might lower barriers to the development of AI tools, and in that way support broadly shared benefits by facilitating greater competition and mitigating concentration of power. On the other hand, just as concentration of technology industries is already a significant challenge, AI might not look much different, and the benefits of this resource may still simply go to a few large firms in “winner takes all-or-most” markets. The workshops discussed how, for instance, large commercial users might be expected to contribute to a fund that supported contributors of training data, or more generally to fund writers, to ensure everyone contributing to the development of AI benefits.

³⁹ For examples of gated access to AI models, see <https://huggingface.co/docs/hub/en/models-gated>.

⁴⁰ As an analogy, consider for instance Wikimedia Enterprise, which “build[s] services for high-volume commercial reusers of Wikimedia content” and charges for that access. https://meta.wikimedia.org/wiki/Wikimedia_Enterprise.

What dataset management practices are necessary?

No matter how a books data commons gets built, it will be important to consider broader aspects of data governance. For example:

- **Dataset documentation and transparency:** Transparent documentation is important for any dataset used for AI training. A datasheet is a standardized form of documentation that includes information about provenance and composition of data, and includes information on management practices, recommended uses or collection process.
- **Quality assurance:** Above, we note the many features that make books useful for AI training, as compared with web data, for example. That said, the institution managing a books commons dataset may still want to collect and curate the collection to meet the particular purposes of its users. For instance, it may want to take steps to mitigate biases inherent in the dataset, by ensuring books are representative of a variety of languages and geographies.
- **Understanding uses:** The institution managing a books commons dataset could measure and study how the dataset is used, to inform future improvements. Such monitoring may also enable accountability measures with respect to uses of the dataset. Introducing community norms for disclosing datasets used in AI training and other forms of AI research would facilitate such monitoring.
- **Governance mechanisms:** In determining matters like acceptable and ethical use, the fundamental question is “who decides.” While this might be settled simply by whoever sets up and operates the dataset and related infrastructure, participatory mechanisms – such as advisory bodies bringing together a broad range of users and stakeholders of a collection – could also be incorporated.

7. Conclusion

This paper is a snapshot of an idea that is as underexplored as it is rooted in decades of existing work. The concept of mass digitization of books, including to support text and data mining, of which AI is a subset, is not new. But AI training is newly of the zeitgeist, and its transformative use makes questions about how we digitize, preserve, and make accessible knowledge and cultural heritage salient in a distinct way.

As such, efforts to build a books data commons need not start from scratch; there is much to glean from studying and engaging existing and previous efforts. Those learnings might inform substantive decisions about how to build a books data commons for AI training. For instance, looking at the design decisions of HathiTrust may inform how the technical infrastructure and data management practices for AI training might be designed, as well as how to address challenges to building a comprehensive, diverse, and useful corpus. In addition, learnings might inform the process by which we get to a books data commons – for example, illustrating ways to attend to the interests of those likely to be impacted by the dataset’s development.⁴¹


While this paper does not prescribe a particular path forward, we do think finding a path (or paths) to extend access to books for AI training is critical. In the status quo, large swaths of knowledge contained in books are effectively locked up and inaccessible to most everyone. Google is an exception – it can reap the benefits of their 40 million books dataset for research, development, and deployment of AI models. Large, well-resourced entities could theoretically try to replicate Google’s digitization efforts, although it would be incredibly expensive, impractical, and largely duplicative for each entity to individually pursue their own efforts. Even then, it isn’t clear how everyone else – independent researchers, entrepreneurs, and smaller entities – will have access. The controversy around the Books3 dataset discussed at the outset should not, then, be an argument in favor of preserving the status quo. Instead, it should highlight the urgency of building a books data commons to support an AI ecosystem that provides broad benefits beyond the privileged few.

⁴¹ For other existing and past examples, one might look to the work of Europeana, <https://www.europeana.eu/en>, as well as the mountain of commentary on the failed class action settlement between Google, the Authors Guild, and the Association of American Publishers – see e.g. the excellent collection of court filings created by James Grimmelman and colleagues (now archived at the Internet Archive) – <https://web.archive.org/web/20140425012526/http://thepublicindex.org/>. The Settlement expressly would have set up a “Research Corpus” for non-consumptive research. HathiTrust created a Research Center, with the intention of becoming one of the hosts for the “Research Corpus.” The Settlement was criticized and was ultimately rejected by the district court for both substantive reasons (that is, what the settlement would specifically do) and procedural (in the sense of violating class-action law, but also in a broader sense of representing a “backroom deal” without sufficient participation from impacted interests). The Research Corpus was not a core locus of critique, though it did receive concern in terms of providing too much control to Google, for example. Our purpose in mentioning this is not to relitigate the issue, but rather to call out that design decisions of this sort have been considered in the past.

Acknowledgements

Authored by Alek Tarkowski and Paul Keller ([Open Future](#)), Derek Slater and Betsy Masiello ([Proteus Strategies](#)) in collaboration with [Creative Commons](#).

We are grateful to participants in the workshops, including Luis Villa, Tidelift and openml.fyi; Jonathan Band; Peter Brantley, UC Davis; Aaron Gokaslan, Cornell; Lila Bailey, Internet Archive; Jennifer Vinopal, HathiTrust Digital Library; Jennie Rose Halperin, Library Futures/ NYU Engelberg Center, Nicholas P. Garcia, Public Knowledge; Sayeed Choudhury; Erik Stallman, UC Berkeley School of Law. The paper represents the views of the authors, however, and should not be attributed to the workshop as a whole. All mistakes or errors are the authors'.

 This report is published under the terms of the [Creative Commons Attribution License](#).